AUTHORSHIP IDENTIFICATION OF INSTANT MESSAGES

M.A.C. Akmal Jahan[1], M.A.C. Jiffriya[2] and M.N. Nawfan[1]
[1]*Dept. of Mathematical Sciences, South Eastern University of Sri Lanka*
[2]*Dept. of Information Technology, Hardy Advanced Technological Institute*
[*]*akmaljahan@fas.seu.ac.lk*

Authorship attribution is a process in which the author of the given text corpus can be automatically recognized using some techniques. In early days the approach to authorship detection was stylometric which is used to identify the particular author of the printed materials, online texts such as blogs, e-mails, tweets, posts etc. In past years e-mails took a big role in communication. In a vast distribution of social media people spend lot of time in online communication like chatting, which nowadays becomes an easiest and effective communication media among people.The social tool like Facebook, Skype, Google talk and the other instant messaging tools contribute greater role in the real time communication rather than the e-mails.In current era, cybercrimes and security threats become a big issue on the all internet related activities.Even though, instant messaging is highly used as fast and effective communication,it is more vulnerable to several attacks and this issue need to be addressed. So far, standard stylometric features have been used for the authorship detection. However, attempts to this approach are still in beginning.Therefore, this paper produces an alternative way for authorship attribution of instant messages. Here, we have used vector space model using unigram technique. Processed chat data set from individual users in which 2/3 of the datais treated as training set and the remaining set is usedfortesting. Similarity score between training and the testing set have been computed using the given algorithm. From the overall result, 75% of the training corpus shows the maximum similarity score with its testing pair. Moreover, the length of the chat corpus does a significant effect on the similarity score which determine the authorship attribution of the instant messages.

Keywords: Authorship attribution, Unigram, Vector space model